

Towards a Diffusion Based Virtual Subject Generator

Imran Nasim^{1,3}, Adam Nasim^{2,3}

¹IBM UK; ² Quantitative Pharmacology, Merck KGaA, Lausanne, Switzerland ;³ Department of Mathematics, University of Surrey

Abstract: Synthetic tabular data generation is critical for numerous real-world applications. VAE and GAN-based architectures have shown success in generating synthetic tabular data but often struggle with domain-specific challenges and the incorporation of temporal dependencies. Recently, diffusion-based models have been used as an alternative to VAEs and GANs, offering enhanced capabilities for synthesizing heterogeneous features. In this work, we employ the TimeAutoDiff model, which integrates the strengths of VAEs and diffusion models to capture both feature and temporal dependencies in tabular data. By integrating the model within an innovative hyperparameter optimization framework, we evaluate TimeAutoDiff’s performance on a challenging real-world preclinical dataset characterized by sparsity and limited unique samples. Our findings demonstrate significant alignment with temporal trends between real and synthetic data. However, deviations in cancer type classification highlight challenges in latent space separation, likely stemming from the VAE’s difficulty in disentangling class-specific features without explicit conditioning. This limitation affects the tumour growth distributions, underscoring the need for improved latent encoding strategies to enhance conditional generation fidelity.

Background: Tabular data is a widely adopted standardized data format, making the ability to generate highly realistic synthetic tabular data essential for a wide range of real-world applications [2]. Due to the ubiquity of this data format, there has been a significant motivation towards deep learning based approaches to synthesizing fake tabular data [10, 11, 9, 1, 6]. Most of these initial approaches were based on the VAE and GAN neural architectures [7, 4] and achieved impressive results for some domain specific tasks. However, recently developed diffusion-based tabular synthesizers have shown significant promise in various tasks by outperforming GAN-based methods[5]. A key feature not addressed by the aforementioned methods is handling tabular data with temporal structure, as the methods described above are limited to generating tables with rows that are independent and identically distributed (i.i.d.). In this study, we leverage the recently developed TimeAutoDiff model [8], which combines the strengths of VAEs and diffusion models to effectively capture temporal dependencies in tabular data. Furthermore, we integrate TimeAutoDiff with an innovative hyperparameter optimization pipeline to evaluate its performance on a real-world pre-clinical dataset, which presents unique challenges such as sparsity and a limited number of unique samples.

Methods and Datasets: We apply the recently developed TimeAutoDiff model which combines two distinct neural architectures: i) VAE, which is used the processed tabular data with temporal dependence into a latent spaces ii) Latent diffusion model (DDPM), which learns the latent distribution of the projected temporal data. A schematic of the TimeAutoDiff architecture is given in Figure 1. Additionally we utilized the Tree-Structured Parzen Estimator (TPE) sampler for efficient exploration of the hyperparameter space, combined with pruning techniques to terminate underperforming trials early. The data for this study comes from patient derived xenografts (PDX) [3]. In these experiments tumour tissue has been taken from patients and implanted into mice, the tumour volume is measured at discrete time points for each subject. Included are data on 3 anticancer agents: ribociclib, binimetinib, and buparlisib. The dosing regimens were 250 mg once a day for ribociclib and 35 mg once a day for buparlisib. Binimetinib was administered twice daily: 228 mice received a dose of 10 mg, and 8 received 8 mg. Each subject was administered with only one single type of drug. Data were available for 5 cancer types for these molecules: breast cancer (BRCA), colorectal carcinoma (CRC), gastric cancer (GC), non-small cell lung cancer (NSCLC), and pancreatic ductal adenocarcinoma (PDAC). The dataset is inherently sparse and has varying sequence length, a common structure within preclinical xenograft datasets. In order for the data to be compatible with TimeAutoDiff we linearly interpolate the data to get measurements for everyday and consider all of the data which sequences with 40 days of measurements $T \in [0, 40]$. The analysis includes 359 mice with an average of 16 observations per mouse.

Results: To evaluate the performance of the model we performed inference on the test data after model training, we plot the feature distributions for the test data in Figure 2. For the tumour size distribution, left panel, which is obtained through flattening all trajectories we observe that the synthetically generated values peak at a smaller tumour size compared to the real data but there appears to be agreement at right tail of the distribution at the larger tumour size values as evident by the overlapping regions. In the case of Drug, Dose and Schedule features (right three panels) we observe a good agreement between the real and synthetic distributions with minor differences seen in classifications. However this difference becomes significant when we consider the cancer type (CTYPE) where we observe a bias in the preference for synthetic generations to be classified as cancer type ‘GC’. We believe this is due to the VAE failing to appropriately separate out the cancer type classes within the latent space. Further, we proceeded to check the agreement between the synthetic and real tumour sizes as a function of time by plotting a visual predictive check (VPC), see Figure 3 (right panel). We observe a good agreement between the synthetic data and real values in particular at the 5th percentile though we find the model tends to overestimate the 95th percentile of the tumour size particularly for earlier times. We plot five representative tumour growth curves for the same subject IDs in Figure 3 (left panel) to check the fidelity of the synthetic data generation process. We find a qualitative agreement where both the synthetic and real curves grow over time but we note a significant spread in growth trajectories for the same IDs.

Conclusion: In this paper we have employed the recently developed TimeAutoDiff architecture to synthetically generate temporally dependent virtual subject data for a real world tabular dataset which presents unique challenges of samples size and sparsity. Though we find a good agreement with a number of covariates distributions (Drug, Dose and Schedule) we find significant deviation in the cancer type classification. We believe this issue likely arises from the VAE’s inability to effectively separate the cancer type classes within the latent space. This behavior is well-known, as VAEs often struggle to separate classes in the latent space without explicit conditioning on class labels or structured enforcement. Additionally, we believe this is a significant factor in the observed difference in tumour growth distribution as an incorrect encoding will lead to an incorrect conditional generation. In future work, we aim to explore modifications to the neural architecture to better handle sparse temporal data and address the disentanglement challenges in the encoding process, ultimately improving the quality of the synthetic data. Finally, we would also like to extend our optimization framework to consider additional objective functions.

Figures:

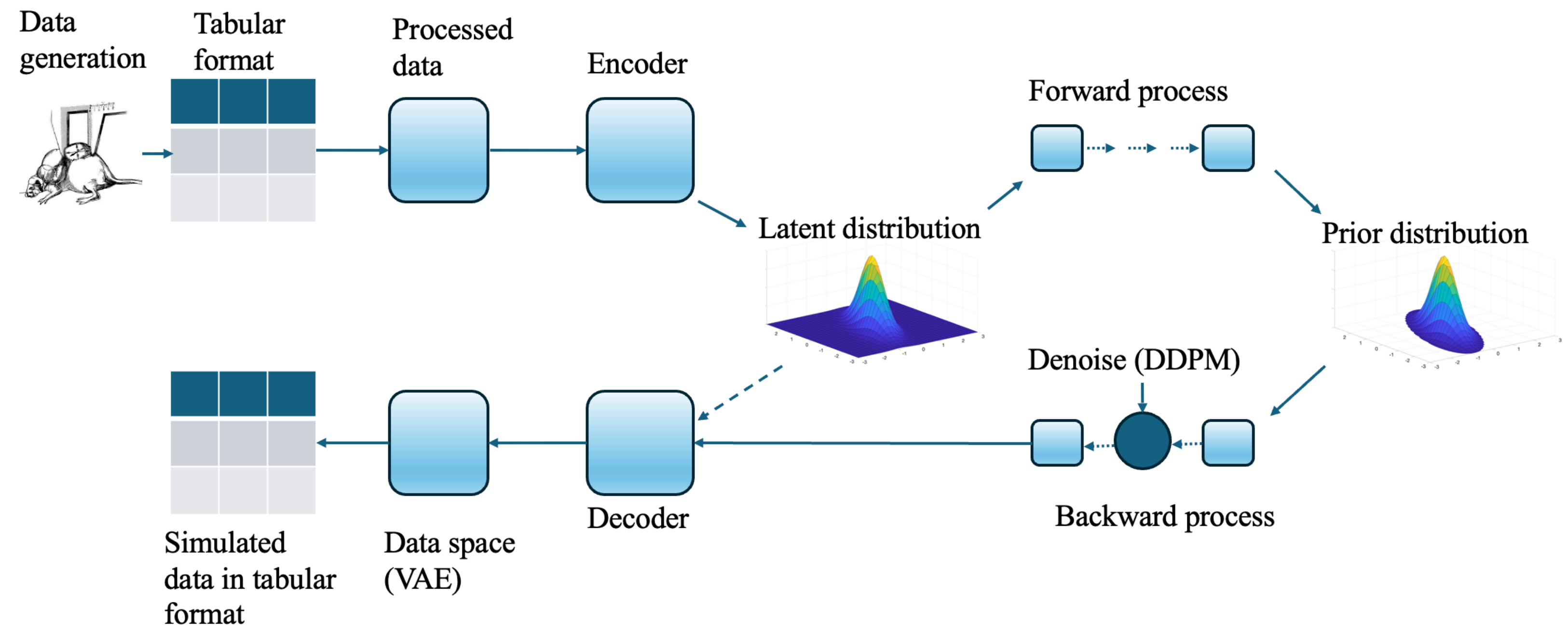


Figure 1: Schematic of the TimeAutoDiff processing steps.

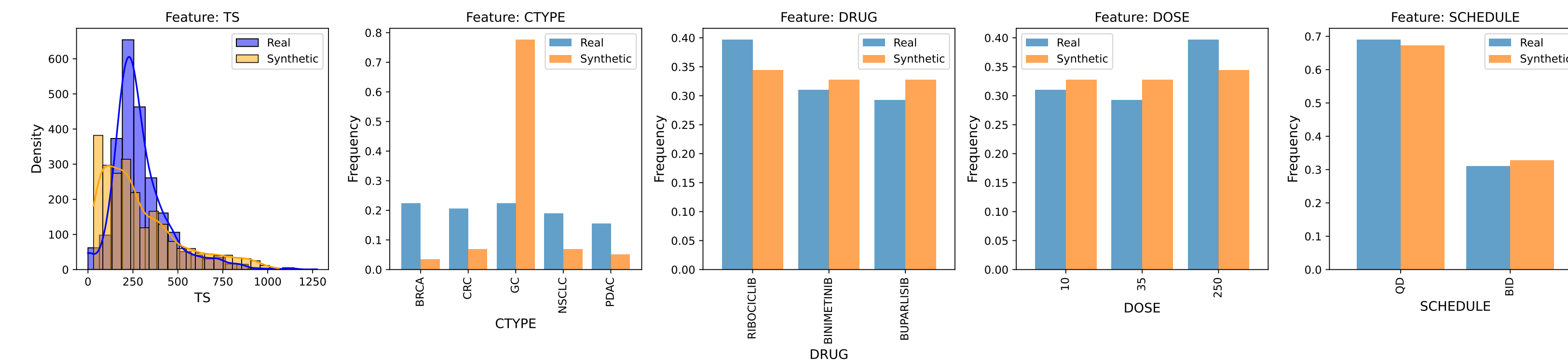


Figure 2: Distribution of features, real vs synthetic, for the test dataset.

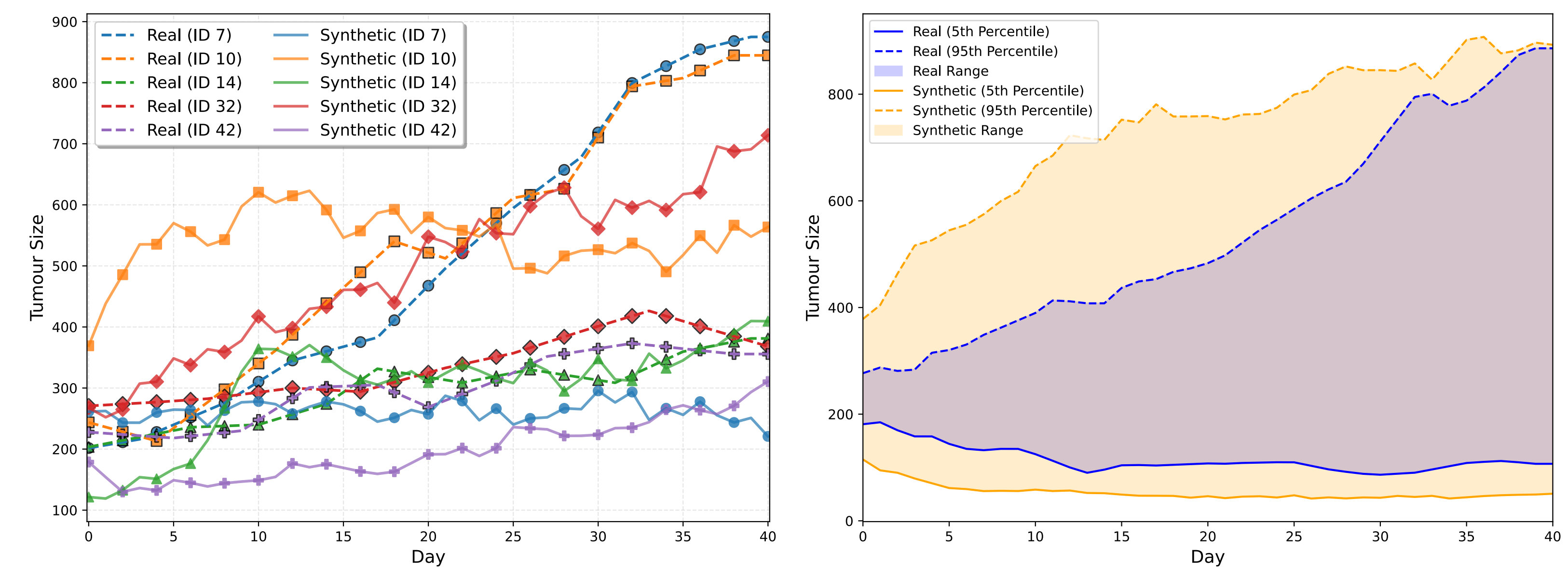


Figure 3: Five representative tumour growth curves for both real and synthetically generated data (left panel). Five representative tumour growth curves for both real and synthetically generated data (right panel).

References

- [1] H. Akrami et al. “Robust variational autoencoder for tabular data with beta divergence”. In: *arXiv preprint arXiv:2006.08204* (2020).
- [2] J. Fonseca and F. Bacao. “Tabular and latent space synthetic data generation: a literature review”. In: *Journal of Big Data* 10.1 (2023), p. 115.
- [3] H. Gao et al. “High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response”. In: *Nat. Med.* (2015).
- [4] I. Goodfellow et al. “Generative adversarial networks”. In: *Communications of the ACM* 63.11 (2020), pp. 139–144.
- [5] J. Kim, C. Lee, and N. Park. “Stasy: Score-based tabular data synthesis”. In: *arXiv preprint arXiv:2210.04018* (2022).
- [6] J. Kim et al. “Oct-gan: Neural ode-based conditional tabular gans”. In: *Proceedings of the Web Conference 2021*. 2021, pp. 1506–1515.
- [7] D. P. Kingma. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).
- [8] N. Suh et al. “TimeAutoDiff: Combining Autoencoder and Diffusion model for time series tabular data synthesizing”. In: *arXiv preprint arXiv:2406.16028* (2024).
- [9] L. V. H. Vardhan and S. Kok. “Synthetic tabular data generation with oblivious variational autoencoders: alleviating the paucity of personal tabular data for open research”. In: *Proceedings of the 37th International conference on machine learning, ICML HSYS Workshop 2020*. 2020.
- [10] L. Xu et al. “Modeling tabular data using conditional gan”. In: *Advances in neural information processing systems* 32 (2019).
- [11] J. Yoon, D. Jarrett, and M. Van der Schaar. “Time-series generative adversarial networks”. In: *Advances in neural information processing systems* 32 (2019).